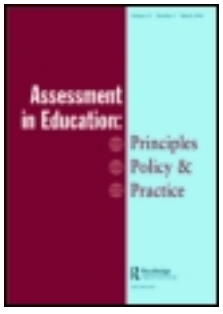

Assessment in Education: Principles, Policy & Practice
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/caie20

Teachers' grading practices: meaning and values assigned

Youyi Sun [a b] & Liying Cheng [a]
[a] Faculty of Education, Queen's University, Kingston, ON, Canada
[b] Department of Foreign Languages, Shanghai Finance University, Shanghai, China
Version of record first published: 04 Mar 2013.

To cite this article: Youyi Sun & Liying Cheng (2013): Teachers' grading practices: meaning and values assigned, Assessment in Education: Principles, Policy & Practice, DOI:10.1080/0969594X.2013.768207

To link to this article: http://dx.doi.org/10.1080/0969594X.2013.768207

Assessment in Education: Principles, Policy & Practice, 2013
http://dx.doi.org/10.1080/0969594X.2013.768207

Routledge
Taylor & Francis Group

Teachers' grading practices: meaning and values assigned

Youyi Sun[a,b]* and Liying Cheng[a]

[a]*Faculty of Education, Queen's University, Kingston, ON, Canada;* [b]*Department of Foreign Languages, Shanghai Finance University, Shanghai, China*

(*Received 9 July 2012; final version received 15 January 2013*)

This study explores the meaning Chinese secondary school English language teachers associate with the grades they assign to their students, and the value judgements they make in grading. A questionnaire was issued to 350 junior and senior school English language teachers in China. The questionnaire data were analysed both quantitatively and qualitatively using Messick's validity framework. Results of these analyses demonstrate that the meaning of the construct *grade* is closely related to two concepts: (1) judgement of students' work in terms of effort, fulfilment of requirement, and quality; and (2) judgement of students' learning in terms of academic enablers (i.e. non-achievement factors such as habit, attitude and motivation that are deemed important for students' ultimate achievement), improvement, learning process, as well as achievement. Two themes concerning these teachers' values in grading were identified: appeal to what is fair and appeal to what is beneficial for students. The teachers were primarily concerned about the consequences of grading on student schooling. These findings shed light on understanding the validity of teachers' grading in the Chinese context where non-achievement factors are valued.

Keywords: grading; meaning; value; validity

Grading is a complex decision-making process that requires teachers to make value judgements as to student learning, achievement, and growth. Studies have shown that teachers tend to consider a hodgepodge of factors such as effort, work habits as well as achievement when assigning grades (Guskey, 2011; Randall & Engelhard, 2009; Yesbeck, 2011). This is discrepant with the suggestion by the measurement community that grades should be based on students' academic achievement or their mastery of learning standards without including confounding factors such as effort and work habits (Dyrness & Dyrness, 2008; McMillan, 2008; Merwin, 1989; O'Connor, 2007; Wormeli, 2006). Brookhart (1991, 1993) suggests that the discrepancy between recommended and actual grading practices is a symptom of a validity problem that can be best framed by Messick's (1989) conceptualisation of validity. Considering teachers' grading practices in light of Messick's validity framework entails exploring the meaning teachers associate with grades, and the value judgements they make when assigning grades, that is, how teachers define the construct of grade, that is, their interpretation of what a grade represents, how they think about grade use and consequences, and what values they place on grades.

*Corresponding author. Email: youyi.sun@queensu.ca

From a sociocultural perspective (Vygotsky, 1978; Wertsch, 1998), the grading decisions teachers make convey their values, beliefs and assumptions about teaching and learning, which are rooted in and shaped by the sociocultural and historical contexts where instruction takes place. In China, public examinations have played a long, important and yet socially accepted role, and teachers' grades impact students' lives in school and beyond (Cheng, 2010). Therefore, it is particularly important to investigate Chinese teachers' grading practices. However, empirical studies that explore Chinese teachers' grading practices are very limited. Even fewer studies have attempted to examine the meaning teachers associate with grades and the values they place on grades that may drive their grade interpretation and use in this context. This study aims to explore the meaning Chinese secondary school English language teachers associate with grades they assign to their students, and the value judgements they make in grading. Specifically, it addresses the following two research questions:

(1) What meaning is associated with grades assigned by Chinese secondary school English language teachers?
(2) What kinds of value judgements do these teachers make when assigning their grades?

Teachers' grading practices

In the simplest sense, a grade is a type of assessment judgement and 'grading' is the process of assigning a grade. Newton (2007) proposes that grading should not be discussed as a specific assessment purpose at the decision level; rather, it should be considered at the judgement level as purely a standards-referenced technical process. However, as McMillan (2008) pointed out, even when teachers use the same grading scale and the same grading guidelines, there is little consistency in teachers' grading across schools. Teachers vary considerably in terms of weighting different factors in determining grades and unforeseen unique situations constantly arise in the classroom setting, which require teachers to make professional decisions.

There is ample evidence derived from studies on classroom teachers' assessment and grading practices that grading is a complex decision-making process influenced by various internal and external factors. McMillan and Nash (2000) proposed a model of teachers' classroom assessment and grading decision-making including both internal and external influencing factors. The most salient internal factor was the teacher's philosophy of teaching and learning. The major external factors were identified as mandated statewide learning factors and high-stakes tests, district grading polices and parents. McMillan and Nash's (2000) model has been supported by studies conducted in other contexts. For example, Cheng and colleagues (Cheng, Rogers, & Hu, 2004; Cheng, Rogers, & Wang, 2008; Cheng & Wang, 2007) compared classroom assessment practices including grading practices by teachers of English as a second or foreign language (ESL/EFL) in three tertiary institutional contexts: Canada, Hong Kong and China. Their studies showed that these teachers' preferences for different methods of assessment and grading were influenced by the beliefs they held about assessment, their assessment purposes, their teaching experiences and educational training, the nature of their instructional contexts such

as the goal of the programme, class size as well as the dominance of external large-scale high-stakes standardised testing. Zoeckler (2007) examined how American high school English language teachers attempted to arrive at a fair grade while weighting both achievement and non-achievement factors and the role of teachers' expectations. Results of this study indicated that grading was influenced by the local grading systems, teachers' perceptions of student effort, and their concerns for moral development.

McMillan (2008) argued that one of the most difficult issues in grading is how to deal with non-achievement factors such as effort, work habits and motivation. He refers to these factors as academic enablers. Teachers tend to consider these non-achievement factors in grading because they are traits that teachers cultivate and regard as important for students' ultimate achievement. In a questionnaire survey with elementary, middle, and high school teachers in the USA, Randall and Engelhard (2010) found that under most circumstances, these teachers abided by the official grading policy of the participating school district, assigning grades based primarily on achievement. However, in some borderline cases, they rely more heavily on other student characteristics such as motivation, behaviour and effort. Simon, Chitpin, and Yahya (2010) investigated teacher candidates enrolled in the Bachelor of Education programme in a Canadian university. Their study found that student effort, participation, and late or missed assignments were identified as key concerns in the participants' thinking about grading.

Teachers' values and beliefs about teaching/learning and their considerations of the purposes and consequences of grading provide a rationale for their grading decision-making. Brookhart's (2004b) systematic review of literature in classroom assessment found that the practice of educational assessment occurred at the intersection of three practical bases: instruction, classroom management and classroom assessment, and at the intersection of three theoretical bases: psychology, sociology and measurement. She recommended that in order to evaluate the meaning, value, accuracy, and consistency of classroom assessment information, the intersection nature of classroom assessment should be acknowledged. Using Brookhart's (2004b) framework, Simon et al. (2010) found that the pre-service teachers in their study attached greater importance to assessment for classroom management, student motivation, and social justice purposes, than to support learning. In an earlier paper, Bishop (1992) argued that teachers cannot act as judges and coaches at the same time and suggested that teachers should give up the judging role to external assessment and focus on developing mentoring relationships with their students to fully function as coaches.

Including both achievement and non-achievement factors in grading threatens the interpretability of the grades assigned by teachers. Brookhart (1991, 1993) suggested that the discrepancy between recommended and teachers' actual grading practices is the symptom of a validity problem, that is, teachers' considerations of the consequences associated with grades outweigh their considerations of grade interpretations. Thus, Brookhart (1993) used Messick's (1989) validity framework to investigate teachers' grading practices in terms of the meaning they wish to convey and the value judgements they make when assigning grades to their students. This study revealed that grades are primarily interpreted as representation of students' work and that being fair and developing students' self-esteem and good attitudes toward future school work are teachers' major concerns in grading. However, the limitation of Brookhart's (1993) study is that all the participants were

practising teachers enrolled in the Master of Science in Education programme at a university.

Grading issues have been receiving increasing concerns in recent years among researchers and educators in the Chinese context. Many of the studies, however, focused on system-wide evaluation and scoring reforms in large-scale high-stakes testing (e.g. Guo, 2007; Liu, 2007), comparisons of different grading procedures such as percentage grading and letter grading (Liu, 2005), and standards-based grading (e.g. Bian & Shan, 2006). Studies on English language teachers' day-to-day grading practices within the Chinese secondary school classroom context are non-existent. Therefore, it is important to empirically research teachers' grading practices within the Chinese context. Following Brookhart (1993), this study employs Messick's (1989) validity framework to explore the meaning Chinese secondary school English language teachers associate with grades they assign to their students, and the value judgements they make in grading.

Applying Messick's validity framework to grading

Messick's (1989) validity framework provides one of the most comprehensive, in-depth, and elegant discussions of values of score interpretations and consequences of score uses (McNamara, 2006). It details two interconnected facets of the unitary validity concept. One facet is the source of justification of testing, based on either appraisal of evidence or consequence. The other facet is the function of the test score, being either interpretation or use. By crossing these two facets, Messick conceptualised validity with a progressive matrix. Within this matrix, construct validity (CV) involving score interpretation supported by empirical evidence is the starting point for all further validity considerations. The scope expands as one moves from (1) appraisal of evidence for the construct interpretation per se (CV), to (2) appraisal of evidence supportive of test use (relevance and utility), to (3) appraisals of the value consequences of score interpretation (value implication), and finally, to (4) appraisal of the social consequences (SC).

Brookhart (1993) rightly pointed out that, 'applying Messick's progressive categories to grades, one can consider the degree to which appraisal of values and consequences of grades use is part of teachers' reflections about their grading practices' (p. 125). She proposed the following four questions to describe teachers' reflections about the validity of grades within Messick's (1989) framework in four progressive categories:

(1) Construct validity (CV) – What does the grade per se mean?
(2) Relevance/utility (RU) – What does the grade mean when it is assigned to a student?
(3) Value implications (VI) – What does the grade mean when it is assigned to a student, and further of what value is it?
(4) Social consequences (SC) – What does the grade mean when it is assigned to a student, and of what value is it, and even further what will happen because of it?

In this study, we used these four levels of questions to analyse teachers' reflective comments on their grading and examine the extent to which they considered values and consequences of grades.

Method

Participants

All secondary school English language teachers from 20 schools in five school districts of a northern city in China were invited to participate in the study. In total, 350 teachers completed the questionnaire (76.1% females; 23.9% males). Most of the participants were between 26 and 40 years old (26–30 = 23.9%; 31–35 = 31.6%; 36–40 = 21.8%). The participants were teaching in junior ($n = 188$) and senior ($n = 162$) secondary schools, with their teaching experiences ranging from 1 year to 35 years ($M = 12.6$; SD = 7.1). With regard to level of education, 55.3% hold certificate/diploma in teaching English as a foreign language, 38% hold a BA or B.Ed. degree. Among the participants, 33.7% have completed a full course on language assessment, 32.2% have completed a course in which language assessment was one of the topics and 24.4% have no training in language assessment in the course of their academic studies or in-service training. Their mean workload was 9.2 hours per week (SD = 2.8), with a mean class size of 54.5 (SD = 10.1).

Instrument

The instrument used to collect data in this study was a questionnaire designed in English with key terms translated into Chinese to ensure the participants' accurate understanding of the questionnaire items. The questionnaire was designed based on previous research on teachers' assessment and grading practices (Brookhart, 1993; Frary, Cross, & Weber, 1993; McMillan, Myran, & Workman, 2002). The questionnaire included four sections. The first section measures the extent to which teachers consider different factors in assigning grades on a five-point scale. The second section measures the extent to which teachers use different types of assessment methods to determine grades, again using a five-point scale. The third section consists of three classroom-grading scenarios, each followed by three choices indicating what the teacher would do in that situation and an open-ended question of 'Why did you make this choice' to explore the meaning and values associated with grades assigned by teachers within the context of the scenario. The three scenarios are included in Tables 2–4. These scenarios represent common grading contexts for classroom teachers – effort/ability, missing work, and improvement – and have been used and validated in previous studies (Brookhart, 1993; Manke & Loyd, 1990). Participants responded to these scenarios in two ways: first, with multiple-choice responses indicating what they would do in each of the scenarios; second, with constructed responses providing the rationale as to why they grade in a certain fashion. The participants answered these questions either in English or in Chinese. The Chinese responses were translated into English by the researchers. The last section of the questionnaire included eight items gathering demographic data of the participants (gender and age) and background information related to teaching (educational qualification, teaching experience, grade level taught, workload, class size and training in assessment). The questionnaire was piloted with 15 secondary school English language teachers in China and some minor modifications were made based on the teachers' feedback and item analyses.[1] Findings from the first two sections of the questionnaire showed that these teachers of English considered both achievement and non-achievement factors in grading, placing greater weight on non-achievement factors such as effort, homework and study habits, and that they employed multiple

types of assessment methods, including performance and project-based assessment, teacher self-developed assessment, as well as paper and pencil tests for grading. Both internal and external factors such as the grade level teachers taught, the assessment training they had received, and their class size affect different aspects of their grading decision-making. This paper focuses on reporting on the results of the third section – the meaning teachers associate with grades they assign to their students, and the value judgements they make in grading. This section represents a different construct of grading from the previous two sections of the questionnaire, that is, the value judgement of teachers' decision-making through grading. It is through those scenarios that teachers' decision-making was captured and portrayed.

Data collection

Two English inspectors[2] of the city, one for junior secondary school and the other for senior secondary school, helped contact the teachers of English from 20 schools in five school districts in the city and mailed them the questionnaires as well as the letter of information and the consent form of the study for research ethics. Teachers were informed that their participation was entirely voluntary and were assured that their responses would be confidential and that the information they provided would not be used for identification. Participants took approximately 30 minutes to complete the questionnaire in their offices. The completed questionnaires were mailed to the two inspectors and were collected by the researchers. The return rate was 89.2%.

Data analysis

Using Messick's (1989) validity framework, constructed responses were coded in terms of four progressive categories regarding different aspects of validity: CV, RU, VI, and SC. Table 1 explains the four categories of the validity matrix with examples from the data of this study. Given that Messick's framework is a *progressive* matrix, a score of 1–4 (1 = CV, 2 = RU, 3 = VI, 4 = SC) was assigned to each response to each hypothetical grading scenario. These four categories are hierarchical in nature, that is, if the response in one category such as SC was evident, then usually the response also included thinking in the previous three categories up to that point (CV, RU and VI). Thus, a score was assigned to a response according to the highest level of category reflected in it. For example, a response that included consideration of RU usually included consideration of CV. Thus, a score of 2 would be assigned to the response. The constructed responses were scored by the two researchers and disagreements were resolved by discussion. In this article, *scores* will be used to indicate these values assigned to the teachers' responses to the question, 'Why did you make this choice' after each scenario. The scores indicate teachers' considerations in grading in terms of the four categories (CV, RU, VI and SC), both categorically and hierarchically, in Messick's progressive matrix. Each response received one score. An aggregated score was not computed for each teacher because the purpose of this scoring is to categorise the responses for the follow-up chi-square tests and qualitative analysis. *Choices* will be used to refer to the multiple-choice responses, A, B, or C, to the scenarios, which indicate what the teacher would do in each scenario. Descriptive statistics analyses were performed for the choices and the scores. The median and the mode of the scores for each of the scenarios were calculated to show the extent to which the teachers in general

Table 1. Examples of the validity matrix categories and explanations.

Category	Example	Explanation
CV	No pains, no gains. Because she did not make a serious effort in class, her grade will naturally be low	This is a response to a scenario about a student working below ability. This teacher thinks effort is a part of the construct that grades are designed to measure
RU	If a student doesn't have a right attitude in learning, he will never make any progress. Homework assignments can help a student to learn what he has been taught. If he doesn't do homework, he can't master the knowledge	This is a response to a scenario about a student not turning in homework. This teacher thinks attitude and homework are parts of the construct (CV), and is looking for relevance of the evidence for the construct in relation to progress and mastery of knowledge (RU)
VI	In my opinion, the students' grades are based on quizzes, tests, and homework assignments. It is the grading policy. If one student can't complete one of the tasks, we should not ignore it. Instead, we should give him an objective and fair evaluation. If not, it is unfair to the other good students who can be assigned 'C' or 'D' and who can turn in homework on time and in time	This is a response to a scenario about a student not turning in homework. This teachers thinks quizzes, tests, and assignments are all parts of the construct (CV), is considering the relevance of the evidence to the grading policy and in relation to other students (RU), and considering the fairness of grading (VI)
SC	I would give him a B based on his improvement because it is good for his future learning. I believe improvement is more important than a single grade in the exam for student learning. If he gets a B this time, I think he will work harder in the future	This is a response to a scenario about a student not turning in homework. This teacher thinks a grade reflects improvement (CV), is considering the relevance of the evidence to learning (RU), reflecting on her belief (VI), and considering a potential consequence would be continued effort (SC)

considered different aspects of validity in the grading situation. χ^2 tests were conducted for each of the three scenarios to examine whether teachers' grading decision-making as indicated by their choices (A, B or C) in the three scenarios varied by their considerations of different aspects of validity as indicated by their response scores. Qualitative analysis was further performed within each of the four categories (CV, RU, VI and SC) using the constant comparative method (Strauss, 1987), that is, all of the responses with the same score value across all the three scenarios were examined comparatively to identify concepts concerning teachers' interpretations of grades and considerations of grade uses and the associated consequences in the context of this study.

Results

Scenario choices and scores

Results about the scenario choices and scores are reported in Tables 2–4. Each of the tables includes (1) the scenario; (2) the frequency and percentage of each choice; (3) the median and mode of the scores indicating the four categories in

Messick's (1989) framework; (4) the chi square results; and (5) the crosstabulation table from the chi square analysis.

Scenario 1 in Table 2 is about working to ability as defined in Brookhart's (1993) study. The student Wang Hong with higher ability has made minimal effort, but the quality of her work is reasonably good. In this situation, 123 (42.1%) of the teachers would lower her grade (Choice B), and 100 (34.2%) of them would raise her grade (Choice C). The number of teachers who would grade Wang Hong on the basis of her work quality (Choice A) is relatively small ($n = 69$; 23.6%). A total number of 292 constructed responses to this scenario were scored using Messick's (1989) matrix. One hundred and nineteen (40.8%) out of the total responses were assigned 4 (SC), 108 (37%) assigned 2 (RU), and the numbers of responses assigned 1 (CV) and 3 (VI) are 31 (10.6%) and 34 (11.6%) respectively. The chi-square results show a significant association between the teachers' considerations of the four validity aspects and their choices in the scenario, χ^2 (6) = 42.91, $p < .001$. In terms of the effect size, Cramer's $V = .27$, $p < .001$. In this high-ability and low-effort scenario, teachers who would lower the grade tended to provide RU and SC explanations; those who would raise the grade tended toward SC; further, those who would grade on quality gave all different explanations across CV, RU, VI, and SC.

Scenario 2 in Table 3 is about missing work. The student Li Wen has consistently received a D on each of the tests and his grades on the quizzes ranged from 60% (D) to 75% (C). He did not turn in his homework. In this situation, 125 (43.7%) of the teachers would use a straight average and assign Li Wen an F (Choice A). Ninety-two (32.2%) of the teachers would assign Li Wen a D based on

Table 2. Results for scenario 1 (working to ability).

Scenario 1
Wang Hong, one of the students in your class, has high academic ability, as shown by her previous work, test results, reports of other teachers, and your own observations. As you look over her work for the grading period, you realise two things: the quality of her work is above average for the class, but the work does not represent the best that she could do. The effort she has shown has been minimal, but, because of her high ability, her work has been reasonably good. In this situation, you would:

N	%	
69	23.6	A. Grade Wang Hong on the quality of her work in comparison to the class, without being concerned about the amount of work that she has done
123	42.1	B. Lower Wang Hong's grade because she did not make a serious effort in your class; she could have done better work
100	34.2	C. Assign Wang Hong a higher grade to encourage her to work harder

Why did you make this choice? median score = 3; mode score = 4.

$\chi^2 = 42.91$, $df = 6$, $p < .001$; Cramer's $V = .27$, $p < .001$.

	Score				
Choice	1 = CV	2 = RU	3 = VI	4 = SC	Total
A. On quality	17	21	15	16	69
B. Lower grade	9	58	6	50	123
C. Raise grade	5	29	13	53	100
Total	31	108	34	119	292

Table 3. Results for scenario 2 (missing work).

Scenario 2
You are the English teacher of a class with varying ability levels. During this grading period, the students' grades are based on quizzes, tests, and homework assignments. Li Wen has not turned in any homework assignments despite your frequent reminders. His grades on the quizzes have ranged from 60% to 75%, and he received a D on each of the tests. In this situation, you would:

N	%	
125	43.7	A. Assign Li Wen a grade of 0 for the homework assignments and include this in the grade, thus giving him an average of F for the grading period
92	32.2	B. Ignore the missing homework assignments and assign Li Wen a D
69	24.1	C. Ignore the missing homework and assign Li Wen a C

Why did you make this choice? median score = 2; mode score = 2.

$\chi^2 = 19.32$, $df = 6$, $p < .05$; Cramer's $V = .18$, $p = .004$.

	Score				
Choice	1 = CV	2 = RU	3 = VI	4 = SC	Total
A. Assign an F	11	50	27	37	125
B. Assign a D	9	58	8	17	92
C. Assign a C	3	27	14	25	69
Total	23	135	49	79	286

his test grades (Choice B). The number of teachers who would assign Li Wen a relatively higher grade C based on his quiz grades (Choice C) is relatively small ($n = 69$; 24.1%). In terms of the score, about half (47.2%) of the responses ($n = 135$) of the total 286 responses were assigned 2 (RU), 79 (27.6%) assigned 4 (SC), and the numbers of responses assigned 1 (CV) and 3 (VI) are 23 (8.0%) and 49 (17.1%) respectively. The chi-square results show a significant association between the teachers' considerations of the four validity aspects and their choices in the scenario, χ^2 (6) = 19.32, $p < .05$. A lower effect size was found (Cramer's $V = .18$, $p = .004$). In this missing work scenario, irrespective of what grading decisions teachers would make (assigning 0, ignoring the missing work with a D or a C), RU is the explanation teachers tended to focus on.

Scenario 3 is about improvement, where the student Zhang Lin improved from a D on the first exam to a C on the second one. In this scenario, 86.8% of the teachers ($n = 237$) would assign a higher grade B to Zhang Lin, noting his improvement (Choice B). Less than 10% of the teachers ($n = 27$) would use an average grade of C (Choice A), and even fewer ($n = 9$; 3.3%) teachers would grade Zhang Lin on the quality of his work (Choice C). A total number of 273 responses to this scenario were scored using Messick's (1989) matrix. One hundred and thirteen (41.4%) out of the total responses were assigned 4 (SC), 93 (34.1%) assigned 2 (RU), and the numbers of responses assigned 1 (CV) and 3 (VI) are 28 (10.3%) and 39 (14.3%) respectively. The chi square results show a significant association between the teachers' considerations of the four validity aspects and their choices in the scenario, χ^2 (6) = 29.19, $p < .001$, Cramer's $V = .23$, $p < .001$. In this scenario of improvement, a dominant majority of teachers would make a decision to reward the student's improvement and they tended to focus on SC (Table 4).

Table 4. Results for Scenario 3 (improvement).

Scenario 3
You are the English teacher of a class which consists of students with varying ability levels. For this class you give two exams in each term. As you compute Zhang Lin's grade for this term, you see that on the first exam, he obtained a score equivalent to a D and on the second exam, a B. In this situation, you would:

N	%	
27	9.9	A. Assign Zhang Lin an overall grade of C, which is the average of his scores on the two exams
237	86.8	B. Assign Zhang Lin an overall grade of B, noting that there was improvement in his performance
9	3.3	C. Grade Zhang Lin on the quality of his work in comparison to the class, without being concerned about his improvement

Why did you make this choice? median score = 3; mode score = 4.

$\chi^2 = 29.19$, $df = 6$, $p < .001$; Cramer's $V = .23$, $p < .001$.

	Score				
Choice	1 = CV	2 = RU	3 = VI	4 = SC	Total
A. Assign a C	6	7	11	3	27
B. Assign a B	20	84	26	107	237
C. On quality	2	2	2	3	9
Total	28	93	39	113	273

Across the three scenarios, the frequencies of SC and RU are higher than those of CV and VI. All the values of the chi-square statistics are significant, indicating that the teachers' considerations of the four aspects of validity have a significant effect on their grading decision-making. However, the effect sizes across the three scenarios are all relatively low, ranging from .18 to .27.

Substantive comments

All the responses to the question 'Why did you make this choice' that were assigned the same score value (1 = CV, 2 = RU, 3 = VI, 4 = SC) across all the three scenarios were compared qualitatively. There were 82 responses in the CV category, 336 in the RU category, 122 in the VI category, and 311 responses in the SC category. The qualitative analysis of the responses within each category revealed some major concepts concerning the teachers' considerations of the four validity categories: CV, RU, VI, and SC in the context of this study.

Construct validity

The responses that were scored 1 (CV) across the three scenarios reflect simply the meaning of grades that the teachers assign. The most prominent concept revealed by analysis of these responses is that a grade is a form of reward to students for work done. Teachers make decisions about the reward based on three aspects: (1) students' effort; (2) work quality; and (3) fulfilment of requirement. For example, when explaining why he chose to lower Wang Hong's grade in Scenario 1, a

teacher commented: 'The reason is simple. Because she did not make a serious effort in class. Her grade should represent the effort she made'. A teacher graded Zhang Lin in Scenario 3 based on the quality of his work rather than his test scores and emphasised, 'Good test results always belong to the past. We should grade him on the quality of his work because to a student, the most important thing is the quality of homework'. In Scenario 2, a teacher assigned Li Wen an F because of the missing work and explained that, 'Homework assignment is a requirement, and Li Wen did not fulfill this requirement. If Li Wen had turned in homework assignments, he would get a higher grade'.

Another prominent concept that emerged from the responses of the CV category is that grade is a reflection of the academic enablers (McMillan, 2008). These enablers include cognitive and affective-oriented attitudes and behaviours. For example, a teacher chose to assign Li Wen a 0 for the missing homework in Scenario 2, 'because the missing work suggests he does not have a serious attitude toward study, and the grade should reflect a student's study attitude and his attitude toward the subject'. Another teacher who made the same choice in the scenario made a similar comment: 'A student who always doesn't finish his homework shows that his attitude to learning and the habit of learning is not very good'. In Scenario 1, a teacher chose to assign the student a higher grade, 'because Wang Hong has high ability. So she should get good grades, and we can give her high marks'. The rationale a teacher provided for his decision to ignore Li Wen's missing work and assign him a D in Scenario 2 is that 'homework isn't the only or the most important way to evaluate his ability'.

For many teachers, grade has self-referenced meaning. They graded the students based on their improvement/progress or ability. A teacher chose to assign Zhang Lin a C in Scenario 3 based on the average of his two scores on the two exams, but offered this rationale: 'Because on his first exam, he got a D, but on the second exam, he got a B. That means he made great progress'. Some teachers interpreted the grades they assigned as a general evaluation of the students and/or their learning. For example, a teacher chose to assign Zhang Lin a grade based on his average score because 'this will provide a comprehensive evaluation of the student'. Another teacher chose to assign Li Wen a grade of 0 for the missing homework in Scenario 2 because she thought 'it is the learning process that is really important'. Yet other teachers simply interpreted the grade as a calculated score, as demonstrated in a teacher's reflection on his decision to assign Li Wen a D: 'because I think the grade is the total grade including the quizzes, tests, the student's doing activities and assignment'. Only a small number of teachers interpreted the grade in terms of students' achievement. A teacher who assigned Li Wen an F because of the missing homework commented that, 'homework is a good way of checking the knowledge that the students mastered, so we cannot ignore it when we grade the students'.

The above findings are consistent with the results from the first two sections of the questionnaire. For example, the teachers were found to consider both achievement and non-achievement factors in grading, placing greater weight on non-achievement factors such as effort, homework and study habits. Analyses of the first two sections of the questionnaire also showed that both internal and external factors such as the grade level teachers teach, the assessment training they have received, and their class size, affect their considerations of different factors in their grading decision-making. The following results from the qualitative analyses of the teachers'

responses under the RU, VI and SC categories provide further evidence as to why they consider these factors.

Relevance/utility

The responses that were scored 2 (RU) reflect both what the teachers think grades mean and where they would look for confirming evidence of the use of their grades. These responses show that the teachers think that grades are relevant to and useful for a variety of functions. For example, a teacher commented in Scenario 2: 'As a teacher, we'd better give our students many chances to study. Let them feel happy. We should encourage them as much as possible'.

The function of grading that was most often mentioned in the teachers' responses was encouragement. For example, in explaining her decision in Scenario 3, a teacher wrote, 'I want to give more encouragements to the students. Students need more encouragement in their study'. Similarly, another teacher in Scenario 1 commented, 'I think giving encouragement to students is very important during the teaching procedure. Let students feel that they can do it'. Reminder is another grading function that has been mentioned by 28 teachers, for example, 'I want to remind her of the importance of doing her homework'; 'I would lower Wang Hong's grade to remind her to realise what she should do in future'; and 'This is a way, in my opinion, to remind Wang Hong to make a serious effort and try to get Grade A'. Some teachers, particularly those who decided to lower the students' grades, mentioned the function of grading as warning. For example, a teacher who decided to lower Wang Hong's grade in Scenario 1 said:

> To let her know that although she has got a good result this time, she may fail next time if she doesn't study harder. Those who have a better attitude are more likely to do things well. So every student can succeed if he is hard-working.

Value implication

The responses scored 3 (VI) revealed some value statements about grade interpretation in the teachers' comments. Two themes emerged from these statements. One theme is the appeal to what is fair. When explaining why a teacher chose to grade Wang Hong on the quality of her work in Scenario 1, the teacher commented:

> Justice is justice. Although she has high academic ability, her work doesn't represent the best she could do. I won't choose B or C, which just make a lower or higher grade. That's not fair. I just consider the quality, especially after being compared with the other students in the same class.

Within this theme, fair is the word that was used most often. 'Students need fair treatment', one teacher wrote when explaining her choice in Scenario 2. Many teachers also emphasised students' obligations in their comments. For example, when reflecting her grading decision in Scenario 2, a teacher commented 'as a student, he should also do his homework on time. It's his duty'. Similarly, another teacher said, 'As a student, he or she ought to finish the homework on time, but Li Wen didn't manage to do so. I hope Li Wen could think over about it and study harder'.

The other theme is the appeal to what is beneficial for students, which reveals the teachers' beliefs about teaching and learning. There are three subthemes under this theme: (1) teachers should encourage students; (2) teachers should be strict with their students; and (3) learning is a progressive process. *Encouragement* is a response that was used in many comments, particularly when teachers would assign a higher grade to the student. These comments go beyond the relevance of grades to encouraging students as commented under the category of RU and such comments explicitly express the teachers' belief about the value of using grades to encourage students. In a sense, encouragement is reinforced in VI. 'Encouragement is much better than punishment in my opinion'. This is the rationale a teacher provided for her choice in Scenario 2. Also many teachers emphasised that 'we should be strict with the students, especially those who have high ability', and 'being strict with the student is good for him', particularly when teachers decided to lower the student's grade. Further, a teacher made the following comments on students' progress when reflecting on her decision in Scenario 3:

> Nothing is better than improvement Zhang Lin has made. Each teacher must be glad to see the students' improvement, which is the real happiness for the teacher. What's more, the more encouragement the students obtain from the teacher, the more rapid progress the students will make. So we should look at every student from a progressive perspective, even though his current performance is not as good.

Social consequences

The responses scored 4 (SC) reflect the teachers' considerations of the consequences of giving a particular grade. Most of the responses focus on consequences on student schooling. Three categories of school consequences were most often used in the teachers' comments: change in student effort (e.g. 'I want to encourage him and let him realise his improvement. In the future, he would make every effort on study'); continued improvement/progress (e.g. 'if the student has made great progress, we should let him see this so that he can make greater progress'); and change in student attitude (e.g. 'in the present era, all the children have high self-esteem. As teachers we should pay attention to it. I graded her on the quality of her work in comparison to the class in order to let her be of high self-esteem and then she will be more confident').

Some responses referred to consequences on students beyond their schooling. A number of teachers mentioned that assigning students a higher grade may let them feel loved and increase their self-esteem. Six teachers mentioned the relationship between teachers' grading and students' future success, for example:

> Everything and everyone is not perfect, a student included. So as for a teacher, I think I should give him/her a chance and help him or her create his/her confidence. One day, she/he will make a success in her/his life.

A good number of comments referred to consequences of grading on other students, for example, 'because homework is an important part during the grading period. If I ignore his missing homework, other students will follow him. That's not a good example. So I choose to give him an average of F', and 'in this way, I can make every student obey the class rules. And have a good habit of studying'.

Discussion

This study investigated the meaning Chinese secondary school English language teachers associate with grades they assign to their students, the value judgements they make as well as their considerations of consequences in grading. Using Messick's (1989) validity framework, the extent to which these teachers engage in value judgements in their grading practices was examined. The rationales teachers provided for their choices in the three scenarios reflected their considerations of the four aspects of validity (CV, RU, VI, and SC) in a hierarchical and progressive manner. The chi square results indicate that these considerations have significant influences on the teachers' grading decision-making. Descriptive statistics of the scores further suggest that across all the three scenarios, grade use and the associated consequences were primary concerns of these Chinese secondary school English language teachers, especially for Scenario 1 (high-ability and low-effort) and Scenario 3 (improvement). When assigning grades to students with high-ability and low-effort, teachers may raise or lower the grade considering the consequences such as change in student effort and/or attitude. When assigning grades to students who have made improvement, teachers may raise the grade considering the consequences such as continued improvement and/or change in attitude. Related to consequences are teachers' considerations of the RU of grades. For example, in Scenario 3, teachers may choose to lower the grade based on their considerations of the function of grade as a reminder to the student of the importance of completing homework. These findings confirm Brookhart's (1991, 1993) suggestion that validity of classroom teachers' grading should be considered within Messick's validity framework and our study supports her argument that teachers' considerations of the consequences associated with grades outweigh their considerations of grade interpretations.

Results from the qualitative analysis of the teachers' comments in the CV category show that for these Chinese teachers, the meaning of their grades (the construct of grade) is closely related to two concepts: (1) judgement of students' work in terms of effort, fulfilment of requirement, and quality; and (2) judgement of students' learning in terms of academic enablers, improvement, learning process, as well as achievement. 'Thus, achievement is part of the construct, but not the whole of it' (Brookhart, 1993, p. 139). Our finding is also consistent with the conclusion of previous studies that teachers consider a variety of factors in their grading (Guskey, 2011; Randall & Engelhard, 2009; Yesbeck, 2011). However, in the context of this study, particular weighting seems to be given to effort. Such a finding is not surprising within the Chinese context, where learning is believed to depend upon effort rather than ability (Wang, 2008).

The analysis of value implications (VI) demonstrated in the teachers' reflections on their grading choices revealed two themes: the appeal to what is fair and the appeal to what is beneficial for students. These two themes are closely related to the dual roles teachers play: judges and coaches (Bishop, 1992; Wilson, 1996). As judges, they appeal to what is fair, giving priority to fairness, justice and objectiveness of their judgement-making in the grading practice. As coaches, mentors, or advocates for students, they appeal to what is beneficial for students, building in considerations of encouragement, strictness and improvement/progress through grading. The teachers' choices in the three scenarios indicate that in their grading practice, they mix these two types of roles differently for different students. For

example, while in Wang Hong's scenario, where a high-ability student made minimal effort, the largest number of teachers would lower her grade, in Zhang Lin's scenario, most of the teachers would assign him a higher grade because of the improvement. Encouragement and improvement/progress have also been identified as important considerations in teachers' grading in other contexts (e.g. Brookhart, 1993); however, strictness may be a particularly important consideration in the Chinese context under study here. These considerations are a reflection of the assessment purposes in this context as suggested by Cheng et al. (2004, 2008). It is important to note that these VI also contribute to the meaning the teachers associated with grades. Appeal to what is fair is related to the concept that a grade is a form of reward to students for work done while appeal to what is beneficial for students is related to the concept that a grade is a reflection of the academic enablers and has self-referenced meanings.

The analyses of the VI and SC demonstrated in the teachers' reflections on their grading choices show that students' learning and development are the major concerns of the Chinese secondary English language teachers in this study. This again reflects the value associated with the appeal to what is beneficial for students. The specific categories of the consequences, such as a change in effort, continued progress, and a change in attitude, perfectly map onto the CV of grades. Therefore, in teachers' grading, the distinction between grade interpretation and grade use is blurred (Brookhart, 1993), and grading is not a value-free technical process of assessment as proposed by Newton (2007).

To highlight, findings of this study are consistent with results from previous studies on teachers' grading practices in other contexts. However, teachers in this study attach more importance to effort, encouragement and strictness in their grading. This reflects the significant pastoral role teachers assume in the Chinese learning culture in contrast to the more task-oriented instructional role of the teachers in the North American instructional context (Biggs & Watkins, 2001). In China, the teacher's role about educating students as an all-rounded person is more emphasised, especially at the K-12 school level. A Chinese teacher is often deemed as a mentor, a role model, an authority and a parent who takes care of intellectual, emotional, and affectionate needs of the students (Cortazzi & Jin, 1997).

Conclusion

Grading is one of the most challenging aspects in teaching for teachers to do well (Brookhart, 2004a; Cheng & Wang, 2007). It is a complex and value-laden decision-making process that requires teachers to make professional judgements within the context where instruction takes place. Teachers make these judgements based on the values they hold about their grades, which are in turn associated with the multiple roles they play in the sociocultural context. This study has addressed such complexity using Messick's (1989) validity framework. To further address the discrepancy between the recommended and teachers' actual grading practices and to address the dilemma teachers face in grading, a thorough understanding from the sociocultural perspective of the values teachers hold and the roles they play in particular social, cultural and educational contexts is extremely important.

This study, small in scale and within one particular local context in Northern China, has provided some initial yet important insights into teachers' judgement and decision-making in their grading practice. In order to fully understand such an

important aspect of teachers' assessment practices, larger studies with multi-phased design and studies at different levels of instructional context ought to be conducted. Regardless, the findings from this study add to our understanding of teacher grading and have important implications for investigating the grading practices of teachers of English elsewhere in China and around the world.

Acknowledgements

This study was supported by a SEED research grant from Faculty of Education, Queen's University, Canada. The paper was also supported by an MOE Project of the National Key Centre for Linguistics and Applied Linguistics of Guangdong University of Foreign Studies, China. We thank all the teachers for their participation in the study.

Notes

1. Teachers were asked to review the questionnaire items for clarity, accuracy, and completeness. Appropriate revisions were made with the fine-tuning of the wording and three items were deleted from the questionnaire.
2. English inspectors in China are trained curriculum consultants in the education departments at provincial or municipal levels. They are responsible for assisting teachers and students in their teaching, learning and preparation for the exams, and are also involved in teacher training at different levels.

Notes on contributors

Youyi Sun is a PhD student in the Assessment and Evaluation Group (AEG) at the Faculty of Education, Queen's University and an associate professor in the Department of Foreign Languages, Shanghai Finance University. His primary research interests are the interactions between assessment, teaching and learning, and the consequences and validation of large-scale testing. His recent publications include papers in *The Canadian Journal of Applied Linguistics* and *Assessment Matters*.

Liying Cheng, PhD, is a professor and a director of the Assessment and Evaluation Group (AEG) at the Faculty of Education, Queen's University. Her primary research interests are the impact of large-scale testing on instruction, and the relationships between assessment and instruction. Her publications are in *Language Testing, Language Assessment Quarterly, Assessment in Education*, and *Assessment & Evaluation in Higher Education*. Her recent book is *English Language Assessment and the Chinese Learner* (co-edited with A. Curtis, Taylor & Francis, 2010).

References

Bian, Y., & Shan, H. (2006). Standards-based education: Promoting students' learning by applying the marking rule [基于标准的教育:利用评分规则促进学生学习]. *Theory and Practice of Education* [教育理论与实践], *26*, 30–33.

Biggs, J. B., & Watkins, D. (2001). Insights into teaching the Chinese learner. In D. Watkins & J. B. Biggs (Eds.), *Teaching the Chinese learner: Psychological and pedagogical perspectives* (pp. 277–300). Hong Kong: Comparative Education and Research Centre and Australian Council for Educational Research.

Bishop, J. H. (1992). Why US students need incentives to learn. *Educational Leadership, 49*, 15–18.

Brookhart, S. M. (1991). Letter: Grading practices and validity. *Educational Measurement: Issues and Practice, 10*, 35–36.

Brookhart, S. M. (1993). Teachers' grading practices: Meaning and values. *Journal of Educational Measurement, 30*, 123–142.

Brookhart, S. M. (2004a). *Grading*. Upper Saddle River, NJ: Pearson-Merrill-Prentice Hall.

Brookhart, S. M. (2004b). Classroom assessment: Tensions and intersections in theory and practice. *Teachers College Record, 106*, 429–458.
Cheng, L. (2010). The history of examinations: Why, how, what and whom to select? In L. Cheng & A. Curtis (Eds.), *English language assessment and the Chinese learner* (pp. 13–26). New York, NY: Routledge.
Cheng, L., Rogers, T., & Hu, H. (2004). ESL/EFL instructors' classroom assessment practices: Purposes, methods and procedures. *Language Testing, 21*, 359–389.
Cheng, L., Rogers, T., & Wang, X. (2008). Assessment purposes and procedures in ESL/EFL classrooms. *Assessment & Evaluation in Higher Education, 33*, 9–32.
Cheng, L., & Wang, X. (2007). Grading, feedback, and reporting in ESL/EFL classrooms. *Language Assessment Quarterly, 4*, 85–107.
Cortazzi, M., & Jin, L. (1997). Communication for learning across cultures. In D. McNamara & R. Harris (Eds.), *Overseas students in higher education* (pp. 76–90). London: Routledge.
Dyrness, R., & Dyrness, A. (2008). Making the grade in middle school. *Kappa Delta Pi Record, 44*, 114–118.
Frary, R. B., Cross, L. H., & Weber, L. J. (1993). Testing and grading practices and opinions of secondary teachers of academic subjects: Implications for instruction in measurement. *Educational Measurement: Issues and Practice, 12*, 23–30.
Guo, Y. (2007). Introspection of the multi-evaluation system reform of National College Entrance Examination [对我国高考多元评价制度改革的反思]. *Education and Examinations* [教育与考试]*, 4*, 22–26.
Guskey, T. (2011). Five obstacles to grading reform. *Educational Leadership, 69*, 17–21.
Liu, Q. (2007). Thoughts on reform in the evaluation system of higher education institution enrollment and entrance examination ["百分制"向"等级制"变革中的评价理念误区]. *Southeast Academic Research* [江苏教育学院学报*(*社会科学版*)*]*, 4*, 21–25.
Liu, W. (2005). Misconceptions on the reform from hundred-mark system to grading system [高校招生考试评价体系改革的思路]. *Journal of Jiangsu Institute of Education (Social Science)* [东南学术]*, 21*, 22–23.
Manke, M. P., & Loyd, B. H. (1990, April). *A study of teachers' understanding of their grading practices*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.
McMillan, J. (2008). *Assessment essentials for standards-based education* (2nd ed.). Thousand Oaks, CA: Sage.
McMillan, J. H., & Nash, S. (2000). *Teachers' classroom assessment and grading decision making*. Paper presented at the Annual Meeting of the National Council of Measurement in Education, New Orleans.
McMillan, J. H., Myran, S., & Workman, D. (2002). Elementary teachers' classroom assessment and grading practices. *The Journal of Educational Research, 95*, 203–213.
McNamara, T. (2006). Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly, 3*, 31–51.
Merwin, J. C. (1989). Evaluation. In M. C. Reynolds (Ed.), *Knowledge base for the beginning teacher* (pp. 185–192). Oxford: Pergamon Press.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.
Newton, P. E. (2007). Clarifying the purposes of educational assessment. *Assessment in Education: Principles, Policy & Practice, 14*, 149–170.
O'Connor, K. (2007). *A repair kit for grading: 15 fixes for broken grades*. Portland, OR: Educational Testing Service.
Randall, J., & Engelhard, G. (2009). Differences between teachers' grading practices in elementary and middle schools. *Journal of Educational Research, 102*, 175–185.
Randall, J., & Engelhard, G. (2010). Examining the grading practices of teachers. *Teaching and Teacher Education, 26*, 1372–1380.
Simon, M., Chitpin, S., & Yahya, R. (2010). Pre-service teachers' thinking about student assessment issues. *The International Journal of Education, 2*, 1–22.
Strauss, A. L. (1987). *Qualitative analysis for social scientists*. New York, NY: Cambridge University Press.
Vygotsky, L. S. (1978). *Mind in society*. Cambridge, MA: Harvard University Press.

Wang, F. (2008). Motivation and English achievement: An exploratory and confirmatory factor analysis of a new measure for Chinese students of English learning. *North American Journal of Psychology, 10*, 633–646.
Wertsch, J. (1998). *Mind as action*. New York, NY: Oxford University Press.
Wilson, R. J. (1996). *Assessing students in classrooms and schools*. Toronto: Allyn & Bacon.
Wormeli, R. (2006). Accountability: Teaching through assessment and feedback, not grading. *American Secondary Education, 34*, 14–27.
Yesbeck, D. M. (2011). *Grading practices: Teachers' considerations of academic and non-academic factors* (Unpublished doctoral dissertation). Virginia Commonwealth University, Richmond, Virginia.
Zoeckler, L. (2007). Moral aspects of grading: A study of high school English teachers' perceptions. *American Secondary Education, 35*, 83–102.